A facet-theoretical approach to item equivalency

Borg, Ingwer

A Facet-Theoretical Approach to Item Equivalency

INGWER BORG

Abstract. Three notions of item equivalency are distinguished. They correspond to the back-translation approach, the psychometric IRT approach, and the facet-theoretical approach. The latter defines equivalent item as items that answer the same questions. The question, then, is explicated in terms of its design. This yields the item's blueprint. One can extract such blueprints by studying given items, but the result is generally not unique. Nevertheless, it makes it possible to predict empirical regularities for the items and, therefore, tests for equivalency. If the tests fail, however, item non-equivalency is just one possible explanation. Design-equivalency is, on the other hand, a definitional issue, not an empirical one. The enmpirical issue is the design's usefulness for a particular purpose, usually for answering the research question.

1. Definitions of item equivalency

What one ideally wants in cross-cultural surveys are items that are equivalent in the different language versions of the questionnaire. What does that mean? One rather obvious approach to item equivalency is the operational requirement to first translate an item into the other language and then translate this item back into the original language. The backtranslation should be highly similar to the original item. One of the problems with this approach is it merely guarantees a one-to-many mapping of item wordings. That is, there may be more than just one proper translation of the item, even though they all translate back to the original item. For example, I have been told (Hess Medler, 1993) that if one translates the question 'How are you doing these days?' into Spanish, one has two options: one that asks about the respondent's emotional well-being, another that asks

about his or her objective-material well-being. Translated back into English, they both lead to a question similar to the one we started with.

Moreover, there is, of course, always the possibility that translations are difficult or even impossible because the item addresses issues or concepts that have no meaning in the other language or culture. A less dramatic but common case is the challenge of translating a Likert rating scale, where one wonders whether 'strongly agree' is properly translated into German by 'stimme voll und ganz zu'. This brings in a new, and deeper, issue, one that is addressed by the psychometric approach to item equivalency, which requires that "equivalent items ... evoke a specified response, from the set of permissible responses, with the same probability among individuals with equivalent amounts of the characteristic assessed by the item" (Hulin, 1987, p. 123). The extent to which this is true can be checked via (logistic) regression of the observed response scores for an item onto the estimated 'amount of the characteristic' or by simply comparing item statistics (e.g., the mean or the rank order of mean values of a homogeneous battery of items).

Yet, this IRT (item-response theoretical) approach rests on a statistical model, where "one of the critical assumptions ... is that the latent trait space is unidimensional" (Hulin et al., 1982, p. 823). Hence, the assessment of equivalency is conditional to the validity of the assumed model. The issue is addressed in great detail in the IRT literature, and a variety of models have been proposed. All models, however, are dimensional ones. More importantly, the "motive" of the empirical inquiry does not play any role. In that sense, the IRT approach resembles almost all schools of measurement. They conceive of measurement as a process where one first builds an all-purpose measurement "instrument" or "scale". The instrument is put on the shelf, ready for future utilization. There are a number of tomes in which one can look up such instruments and their psychometric properties; one is the three-volume "ZUMA-Skalenhandbuch" (Allmendinger et al., 1983) which is now under revision with an important change of emphasis, i.e. turning it into a handbook of "items" rather than "scales".

Collections of measurement instruments are useful in applied research. They represent, in a sense, an engineering approach, providing "tools" that have been shown to "work". In science, however, one wants more than just predictive validity. One really wants to establish (empirical) laws that relate to theories. Thus, whether implicitly or explicitly, instruments and items in basic science are never formulated in isolation: the researcher formulates items with some lawfulness in mind, a hypothesis that relates observations on these items to other items and to definitional systems. The hypothesis precedes the items and the particular items used in the empirical investigation are almost always only a sample from a huge *universe* of items. What governs the construction or selection of items is the structural hypothesis.

Hence, I propose that an important aspect of item equivalency should be whether corresponding items *both answer the same substantive-scientific question*. Equivalency of items should therefore be considered in the context of *what it is that the researcher wants to know*. In the above example of translating 'How are you doing these days?' into a Spanish language item, for example, one would have to know whether the researcher wants to assess emotional or material well-being, and, indeed, whether assessing this particular issue is important for the hypothesized lawfulness. If this is known, the translations can be checked against this criterion. In fact, we may decide not to translate the item literally but to rewrite it in a particular subcultural jargon. As long as it assesses the particular type of well-being we want to assess, the phrasing of the item does not matter.

Without knowing the intent of the question, translating and back-translating items may only preserve equivalency of words. And items with similar statistical properties, while both satisfying the same formal model, usually ignore the issue of the universe of content and, in any case, the wider structural hypothesis. This amounts to a sterile form-precedes-content approach to item construction, where the formal machinery is guaranteed to generate a battery of one- or multi-dimensional scales simply by trimming content to the

statistical model. In the end, it remains unclear what exactly is being assessed by such items.

Yet, viewing item equivalency from this perspective, one notes immediately that neither statistical models nor linguistic theories nor any other extrinsic scaffolding suffices in providing good translations. What is needed is that the researcher esplicates his or her research question.

2. Explicating the item's blueprint

The above argument may seem artifical and exaggerated, because in most research in the social sciences, the overall research question is stated quite explicitly. However, it is also true that what each individual item is supposed to assess is almost never explicated – except in experimental research! In experiments, the items are experimental conditions which are typically well-designed. Each such condition asks a particular research question and formulates what is to be recorded as an answer.

In survey research, in contrast, items are typically constructed by mixing intuition, factor-analytic thinking, and, possibly, empirical evidence on certain item statistics. Nevertheless, there is always an *implicit* item design. It may have gotten blurred by statistical tinkering with the item pool - such as rephrasing items that are "factorially ambigious" or even eliminating items that are not well-"explained" by the space spanned by the first few principal components - ,but one can often uncover an implicit item design by carefully studying the items with respect to the semantic variables that are systematically varied throughout the items. Such analyses may even help the researcher to come up with items that focus more sharply on what he or she wants to know.

Consider the following case. Bastide & van den Berghe (1957, p. 690) set out "to determine the patterns of race relations in the white middle class of Sao Paulo". They collected empirical data using items that they categorized into four types:

(a) A list of "stereotypes" where the respondent was asked whether he considered Blacks inferior, equal, or superior to Whites in some sense;
(b) Items on social norms of behavior, such as 'Should Whites and Blacks exchange courtesy visits?'
(c) Items on "actual behavior" of the subjects;
(d) Items of "hypothetical personal behavior", such as 'Would you go out with a black person?'

These four types imply, of course, rules for constructing or culling relevant items. However, these rules are not further explicated, and so translating items such as those shown above in (b) and (d) is unnecessarily difficult. What is meant by 'courtesy visits', what behavior does 'go out with' refer to? These are rather obvious problems, but what is more important is that it remains unclear what roles 'courtesy visits' and 'go out with' play in the context of what the researchers sets out to study, i.e. “patterns of race relations”. In other words, is it important that the items contain these semantic elements? As a further example, consider the item 'Are Whites more intelligent than Blacks?'. This seems relatively easy to translate, but is it important that we explicitly refer to 'intelligence'? Or does this item attempt to provide just one piece of evidence in an effort to assess general feelings of superiority of white persons relative to black persons?

Guttman (1959), reanalyzing the Bastide & van den Berghe study, attempted to abstract some of the distinctions made by the item classes (a)-(d). The approach for doing this is actually quite simple. The first step consists of writing each item type as a complete sentence so that the different sentences are structurally as similar as possible. For the Bastide & van den Berghe items, Guttman proposed the following scheme:

1. Belief of a white person that Whites are superior to Blacks on desirable traits.
2. Belief of a white person that Whites should socially interact with Blacks.
3. Belief of a white person that he or she would socially interact with Blacks.
4. Overt action of a white person in the domain of social interactions with Blacks.

The four sentences can be interpreted as rules for allocating given items to a particular item type or as rules for constructing particular types of items. For example, one finds that the item 'Would you go out with a black person?' belongs to class 3.

Figure 1. A mapping sentence for the Bastide & van den Berghe items on patterns of race relations in the white middle class of Sao Paulo.

		A=modality							
	(	a_1 = belief	)						
The	(	a_2 = overt action	)	of a white person (p) vis-à-vis					
	(	a_3 = emotion	)						

	B=relation			*C=reference group*		
(	b_1=comparison	)		(	c_1=Whites in general	)	
(		)	of	(		)	with Blacks
(	b_2=interaction	)		(	c_2=himself/herself	)	

		range		
	(	very positive	)	
➔	(	to	)	towards the relation of the reference
	(	very negative	)	group to Blacks

The next step is to analyze in what ways the four classes of items differ among each other by asking what semantic dimensions are *systematically* varied over the item classes. The semantic material contained in the items that is unsystematic is either likely not to be of direct importance to the scientific question addressed by those who formulated the items or it may reflect an unsystematic item design.

Let us extract what is varied systematically over the four item types. We note, first of all, that the first three item classes assess 'beliefs', the fourth 'overt action'. This constitutes

the first *facet* of the item's blueprint. In facet theory, we write this facet in set-theoretical notation as A = {belief, overt action} and assign to it the name 'modality (of attitudinal behavior)'. Then, we note that the first two item types refer to Whites in general, the latter two to the respondent him/herself. This constitutes facet B = 'reference group' = {Whites, respondent him/herself}. Finally, the first item class assesses comparison behavior, the other item classes refer to interaction behavior. Hence, facet C = 'relation (of respondent to reference group)' = {compare (with respect to desirable traits), interact}. Note that the facets thus extracted reflect a particular perspective, namely the perspective of a psychologist who uses a particular technical language. Notions such as 'belief' or 'overt action' have technical meanings in psychology.

Conceptual clarity can be further enhanced by not only listing the various facets, but by interrelating them within a particular framework, a mapping sentence. This also forces one to explicate the range of the items ("the response scale"). In the given case, one such mapping sentence is shown in figure 1the following one:

The mapping sentence shows that the items of this study all assess attitudes of the respondent towards different forms of behavior of Whites towards Blacks, because they all assess the extent to which a behavior of a reference group is positive or negative towards a common object (Borg & Shye, 1995). Therefore, we may immediately extend facet A to include the usual third "component" of attitudes, i.e., emotions.

Promoting the formality of the item's design reveals, moreover that the item types are not well-designed in one important aspect. Item type 2, by referring to "should" behavior, refers to norms on interracial behavior, while the other items refer to actual behavior or to behavior that is probable. This is a theoretically important distinction, and the translator must know whether it is also important to the researcher. If the original item is vague, and if its measurement intention remains hidden, the translated item is likely to be unclear in the desired research sense. One might decide, therefore, to express the additional distinction just noticed by introducing a fourth facet, 'factuality of the interracial relation' = {certainly exists, presumably exists, is desirable}. The translator, of course, cannot

figure out for him/herself what role this facet plays in the empirical inquiry. It is the task of the researcher to clarify this issue.

Let us now return to the question of what the researcher wants to know. Bastide & van den Berghe wanted to study “patterns of race relations”. Our analyses show that their item typology constrains this research question effectively to a study of patterns of attitudes on race relations. It thereby builds a bridge to what is already known about attitudes in general. Thus one “pattern” hypothesis is that such attitudinal items are positively intercorrelated, reflecting the first law of attitudes (Borg & Shye, 1995). Another “pattern” hypothesis is that the facets built into the items or, expressed differently, projected into these items by a psychologist’s interpretation, are reflected in the structure of the data, in the sense that the items can be statistically discriminated along these facets. One way of testing this discriminability is to ask if the items form non-overlapping regions in a multidimensional scaling representation (Borg & Groenen, 1997).

3. Some comments on mapping sentences

Asking the translator (and not the researcher)to explicate the item’s design is, of course, not the ideal way to proceed. Yet, from experience, I know that this situation is not as unlikely as it may seem. I have been asked on several occasions to provide a facet analysis for a set of items given to me, without being told the purpose of the items in any but an exceedingly vague way. We saw in the above that such a facet analysis is possible. However, it should be obvious that the mapping sentence we came up with is not the only one that is conceivable. Indeed, we pointed out that this particular mapping sentence is one that relates to a psychological background, with technical notions that are obvious only to the psychologist. But even psychologists would, of course, not always arrive at the same facets, because different psychologists operate within different theories.

Borg (1991), for example, classified work value items ('How important is work outcome XYZ for you?', with the range 'not important ... very important') in a variety of different

ways, each reflecting one particular theory. The mapping sentence used in this analysis distinguished two facets of the work outcomes; 'need served by work outcome' and 'performance-dependency of work outcome'. Since various classification schemes for needs exist (e.g., the Maslow hierarchy, Alderfer's ERG theory, and Herzberg's dichotomy), there are also different ways to facetize work values. Similar arguments hold for the second facet. Which of these possible facetizations is to be preferred depends on the purpose of the study. Indeed, depending on the purpose, one may opt for different sets of facets. An analogy in this context is the classification of matter, where the purpose at hand determines whether Mendeleev's periodic table of chemical elements is better than, say, the archaic earth-wind-fire-water distinction.

Apart from the purpose, however, a number of general criteria can be formulated for judging the goodness of a mapping sentence. Since the mapping sentence is a definition and not a hypothesis, "truth" is not an issue. Clearness, however, *is* relevant. Further criteria are its reliability for classifying items, for constructing items, and for communicating about items (among experts). Ideally, a mapping sentence should also be empirically useful. Empirical usefulness is the testable hypothesis associated with the mapping sentence definition. It predicts, among other things, that the conceptual structure induced by the mapping sentence into a pool of items is mirrored in a corresponding structure of the observations.

One cannot expect that a translator by him/herself will, in general, come up with anything else but a mapping sentence that is "superficial", focusing on rather concrete distinctions made by the items and considering their *apparent* purpose only. A "deeper" mapping sentence usually involves considerable expertise. Moreover, good mapping sentences typically develop over time in bidirectional, mutually constraining interaction between conceptual-theoretical work and empirical testing, a cooperative alternation which almost always involves many mapping sentence modifications and item reformulations. Advanced mapping sentences, therefore, become rather abstract and hard to understand for the uninitiated. Translators must ultimately not only be knowledgable about the

languages involved in the translation task, but also at least be able to understand what the researcher wants to know. This requires substantive expertise.

The mapping sentence is the items' blueprint, but this blueprint is often not fully developed before one begins to construct items. Typically, one begins with a vague notion of commonality and then writes down a set of items. One then studies these items – similar to what we did above – to find facets that are likely to make a difference, conceptually or empirically. Then, a first mapping sentence is sketched. This mapping sentence is best tested against new items: They often cannot be reliably classified by the first-draft mapping sentence, and so more conceptual work has to be done on this mapping sentence (sharper definitions, additional facets, better "grammar", etc.). If, after some such iterations, one arrives at a conceptually sufficiently clear mapping sentence, data are collected and the mapping sentence is tested for its empirical usefulness. But the empirical structure of the data may also suggest conceptual structure, as we all know from exploratory data analysis. So the mapping sentence and data related to it are related in some kind of "partnership", i.e. in the basic scientific ping-pong relation of theory and observation.

4. Predictions from facet-designed items and assessing the effects of bad translations

For the Bastide & van den Berghe items, a whole set of predictions can be derived from their mapping sentence. We noted above that one can predict positively intercorrelated items or a regionality of MDS representations that reflects the facets. A more intricate prediction is that the whole set of items forms a system of interrelated cumulative scales, a partial order of Guttman scales (Borg, 1994). These issues are described in detail elsewhere. They are not of particular importance to the topic of this paper, but it should be pointed out that structural hypotheses are an automatic by-product of mapping sentence designs. Hence, mapping sentence designs allow one to check the equivalence of

different-language versions of the items empirically in the usual sense of construct validity.

Several related examples for concrete cross-cultural applications of this approach are the studies by Borg (1986, 1991), Elizur et al. (1991), Borg & Braun (1996). They studied work value items, i.e. items such as 'How important is it to you to make a lot of money in your job?' or 'How important is it to you to have interesting work?'. The mapping sentence for these items distinguished two facets. One facet classified items according to the need that a particular outcome relates to (e.g., in one particular formulation, whether the outcome satisfies an existential-material need, a social-emotional need, or a growth need). The other facet distinguished whether the outcome is performance-dependent or system-dependent. Work value items that were used in surveys conducted in countries such as China, West-Germany, East-Germany, Israel, and the USA were all classified by these facets in the same way. It was found that the data from all countries could be structured equivalently by this facet design. That is, the various items could be statistically discriminated by each facet in turn. Moreover, the pattern of discrimination was the same for each country, i.e., a so-called radex structure in two-dimensional MDS space. No further detail on what exactly this means is needed to see that the data analysis is driven by the content facets, not by a preconceived formal notion such as unidimensionality as in ICC. Indeed, even a multidimensional analysis that concentrates on interpreting dimensions of the items (such as factor analysis) would not have revealed the facets' roles in the data (Borg & Groenen, in press).

Yet, this leads to the question how one should evaluate the situation if no such structural similarities are found. In particular, is it possible to separate effects of bad translation from other effects, such as systematic differences of the samples or non-validity of the design facets in certain cultures? It seems to me that this is not possible. That is, only if structural similarity is given, may one then conclude that the items are equivalent, too. Otherwise, it is a challenging task to disentangle the confounding of effects. An easy solution is to eliminate the problem *by assumption*. This is what is done in traditional

psychometrics: if the underlying true structure of the construct assessed by the items is assumed to be identical, and if the samples are assumed to be homogeneous, then any systematic differences in item statistics that remain after admissible fitting transformations are due to bad translation.

5. Establishing item equivalency independently of the research question?

The equivalency issue is not confined to cross-cultural research. It is equally relevant when one wants to construct parallel tests, for example, or when one considers replications with "similar" items. In a sense, even replicating an empirical investigation with the same items may raise the equivalency issue. In the end, it is not difficult to see that it is a fallacy to believe that one may be able to resolve the issue independent of the research problem, by first establishing instruments with equivalent items before turning to the research question one is really interested in.

The main reason is that in the traditional psychometric approach, items are first selected on the basis of a substantive rule. They are then studied empirically for certain formal properties, in particular for their dimensional structure. Items that do not fit this structure are eliminated or rewritten. However, this also affects, indirectly, the initial rule for constructing or selecting items. One cannot, for example, first pick items because they belong to the domain of attitude items on “race relations in the white middle class of Sao Paulo”, and then decide on statistical grounds that some of them have to be eliminated afterwards. What belongs or does not belong to a universe of items is not a statistical but a definitional issue. The data only reveal the structure of this domain, but cannot affect its content

Making the common blueprint of the items as clear as possible, establishes *one* feature of equivalency, i.e., *design equivalency*. It makes it possible to map items into *design-equivalent* items rather than to translate them literally and trying to preserve irrelevant or

even distractive semantic material. Further advantages of this approach are that it facilitates the identification of item types; it helps modeling the conceptual structure of items; it systematically lays out the universe of items, not just ad-hoc collections of items with a vague notion of communality; it suggests structural laws; and it thus enables one to see common content-driven (not statistically induced) structure in one or several sets of items.

Other approaches to item equivalency may or may not be compatible with or complementary to the facet-theoretical approach. The psychometric method, obviously, does not belong to this set. Both methods are, in a sense, opposites of each other, one starting with content and then proceeding to data and models, one starting with models and fitting content to the models.

An obvious special case of the facet approach is the MTMM approach. 'Method' and 'trait' are just two facets that distinguish among different items. Usually, the MTMM approach is also special in a statistical sense, because researchers who use MTMM these days also use particular (usually linear) statistical methods to analyze the data. Yet, there is no compelling reasons for combining the MTMM approach with such statistical models. In fact, the original work by Campbell & Fiske (1959) looked for certain patterns in the MTMM matrix rather than attempting to fit a particular statistical model. Borg & Groenen (19##) showed, moreover, that the traditional models may be easily replaced with the usual content-driven techniques such a regional MDS, where the regions, of course, relate to 'method' and to 'trait'.

References

Allmendinger, J., Schmidt, P. & Wegener, B. (eds.), *ZUMA-Handbuch Sozialwissenschaftlicher Skalen*. Mannheim und Bonn: ZUMA und IZ.

Bastide, R. & van den Berghe, P. (1957). Stereotypes, norms and interracial behavior. *American Sociological Review*, *22*, 689-694.

Borg, I. (1986). A cross-cultural replication on Elizur's facets of work values. *Multivariate Behavioral Research*, *21*, 401-410.

Borg, I. (1991). Multiple facetizations of work values. *Applied Psychology: An International Review*, *39*, 401-412.

Borg, I. (1994). Evolving notions of facet theory. In I. Borg & P.Ph. Mohler (Hrsg.), *Trends and Perspectives in Empirical Social Research* (178-200). New York: DeGruyter.

Borg, I. & Groenen, P.F.J. (1997). *Modern Multidimensional Scaling*. New York: Springer.

Borg, I. & Groenen, P.F.J. (in press). Regional interpretations in MDS. In M. Greenacre & J. Blasius (eds.). *Visualizing Categorical Data*. New York: Academic Press.

Borg, I. & Shye, S. (1995). *Facet Theory: Form and Content*. Advanced Quantitative Methods in the Social Sciences, Vol. 5. Newbury Park, CA: Sage.

Elizur, D., Borg, I., Hunt, R. & Magyari-Beck, I. (1991). The structure of work values: a cross cultural comparison. *Journal of Organizational Behavior*, *12*, 21-38.

Guttman, L. (1959). A structural theory for intergroup beliefs and action. *American Sociological Review*, *24*, 318-328.

Guttman, L. (1971). Measurement as structural theory. *Psychometrika*, *36*, 329-347.

Hess Medler, S. (1993). Personal communication at the Fourth International Facet Theory Conference. August. Prague, Czech Republic.

Hulin, (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of cross-cultural psychology*, *18*, 115-142.

Hulin, C.L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, *67*, 818-825.